Interprétation spectrale de la classification relationnelle

Lazhar Labiod, Younès Bennani

LIPN UMR 7030, Université Paris 13
99, avenue Jean-Baptiste Clément, 93430 Villetaneuse
Prénom.Nom@lipn.univ-paris13.fr,

Résumé. Ce papier présente une vue spectrale sur l'approche de l'analyse relationnelle pour la classification des données catégorielles. Il établit d'abord le lien théorique entre l'approche de l'analyse relationnelle et le problème de classification spectrale. En particulier, le problème de classification relationnelle est présenté comme un problème de maximisation de trace, ce problème est donc transformé par la relaxation spectrale en un problème d'optimisation sous contraintes qui peut être résolu par des multiplicateurs de Lagrange, la solution est donnée par un problème de valeurs propres.

1 Introduction

Le Clustering a reçu une attention accrue comme un problème important en apprentissage non supervisé avec de nombreuses applications, un certain nombre de différents algorithmes et des méthodes ont émergé au fil des années. La plupart des techniques utilisent une mesure de similarité par paire pour mesurer la distance entre deux points de données. Récemment, des recherches ont été élaborées sur les approches de clustering de données catégorielles (Li et al,2004) (Huang, 1998), où les données catégorielles sont celles dont les valeurs d'attributs sont nominales et non ordonnées, par exemple, la couleur et le sexe. Il est à noter que les difficultés se posent en matière de clustering des données catégorielles en raison de l'absence de liens intrinsèquement ordonnés de données catégorielles. La plupart des techniques de clustering basées sur les mesures métriques de distance ne sont donc pas applicables à la classification des données catégorielles. D'autre part, les méthodes spectrales ont été utilisées efficacement pour résoudre un certain nombre de problèmes de partitionnement de graphes, les méthodes de coupure minimale Ratio-cut et N-cut ont été utiles dans de nombreux domaines, tels que l'agencement du circuit et la segmentation d'image (Shi et al,2000). Étant donné un ensemble de données, la minimisation des fonctions objectives des méthodes N-cut et Ratio-cut est un problème d'optimisation NP-difficile. Ainsi, en général ce problème est transformé par la relaxation spectrale en un problème d'optimisation avec une contrainte qui peut être résolu par des multiplicateurs de Lagrange, et la solution est donnée par un problème de valeurs propres. L'approche de l'analyse relationnelle a été utilisée pour la classification de données catégorielles (Marcotorchino et al, 1978) (Marcotorchino, 2006). Dans ce papier, nous montrons que le problème de la classification relationnelle peut être formellement modélisé comme un problème de maximisation de trace. Nous avons également établi le lien entre la méthode de classification spectrale et l'Analyse relationnelle (AR) qui est fondée sur le critère

de Condorcet. Nous développons ensuite une procédure spectrale efficace inspirée de l'algorithme proposé par (Ng et al, 2001) pour trouver la partition optimale maximisant le critère de l'AR. Les résultats expérimentaux montrent l'efficacité de notre approche.
Le reste du papier est organisé comme suit: La section 2 introduit l'approche de l'analyse relationnelle. La section 3 présente le critère de Condorcet normalisé et son équivalence avec les critères inertiels. Des discussions sur la connexion spectrale de la classification relationnelle et la procédure d'optimisation proposée sont données dans la section 4. La section 5 montre nos résultats expérimentaux et enfin, la section 6 présente des conclusions et certains travaux futurs.

2 Analyse Relationnelle

L'Analyse Relationnelle à été développée en 1977 par F. Marcotorchino et P. Michaud, et s'inspire des travaux de Marquis de Condorcet, qui s'est intéressé au 18ème siècle au résultat collectif d'un vote à partir de votes individuels. Cette méthodologie est basée sur la représentation relationnelle (comparaison par paires) des différentes variables et l'optimisation sous contraintes du critère de Condorcet.

2.1 Notation et définitions

Soit I un ensemble de données avec N objets $I = \{O_1,O_2,...,O_N\}$ décrit par l'ensemble V de M attributs (ou variables catégorielles) $\{V^1,V^2,..,V^m,...,V^M\}$ chacun ayant $p_1,..,p_m,...,p_M$ catégories, respectivement, et soit $P = \sum_{m=1}^{M} p_m$, désigne le nombre total de catégories de toutes les variables. Chaque variable catégorielle peut être décomposée en une collection de variables indicatrices. Pour chaque variable V^m, les p_m valeurs correspondent naturellement aux nombres de 1 à p_m et $V_1^m,V_2^m,...,V_{p_m}^m$ sont des variables binaires tels que, pour chaque $j, 1 \leq j \leq p_m, V_j^m = 1$ si et seulement si V^m prend la $jieme$ valeur. Ainsi la matrice de données peut être exprimée comme une collection de M matrices K^m, $(m = 1,..,M)$ de terme général k_{ij}^m tel que : $k_{ij}^m = 1$ si l'objet i possède la categorie j de V^m et 0 sinon, ce qui donne la matrcie disjonctive K de dimensions $(N \times P)$; $K = (K^1|K^2|...|K^m|...|K^M)$. La matrice disjonctive pondérée $\tilde{K}$ est obtenue par la multiplication de chaque entrée k_{ij} du K par la racine carrée du produit des sommes marginales de la ligne i, $ki.$ et la colonne j, $k.j$, Autrement dit, chaque entrée $\tilde{k}_{ij} = \frac{k_{ij}}{\sqrt{k_{i.}k_{.j}}}$. En terme de notations matricielles on écrit : $\tilde{K} = R^{-\frac{1}{2}}KC^{-\frac{1}{2}}$ où $R = diag(Ke)$ et $C = diag(K^t)e$, avec $e = \mathbf{1}$ est le vecteur de dimension appropriée dont toutes ses valeurs valent 1 et $diag(.)$ est la matrice diagonale. La matrice $\tilde{S}$ est appelée la matrice de Condorcet de terme général $\tilde{s}_{ii'}$, représentant la mesure de similarité globale entre les deux objets O_i et $O_{i'}$, mesurée sur tous les M attributs. La matrice $\overline{\tilde{S}}$ de terme général $\bar{\tilde{s}}_{ii'}$ représente la mesure de dissimilarité globale entre ces deux objets. Pour obtenir la matrice $\tilde{S}$, chaque attribut V^m est transformé en une matrice carrée $\tilde{S}^m = \tilde{K}^m(\tilde{K}^m)^t$ de taille $N \times N$ et de terme général $\tilde{s}_{ii'}^m$, représentant la mesure de similarité entre deux objets O_i et $O_{i'}$ pour l'attribut (variable) V^m. Pour obtenir la matrice $\overline{\tilde{S}}$, une mesure de dissimilarité $\bar{\tilde{s}}_{ii'}^m$ entre les objets O_i et $O_{i'}$ pour l'attribut V^m est alors calculée comme le complément à la mesure de similarité maximale possible entre ces deux objets. Comme la similarité entre deux objets différents est inférieure ou égale à leur auto-similarité : c.a.d $\tilde{s}_{ii'}^m \leq min(\tilde{s}_{ii}^m,\tilde{s}_{i'i'}^m)$, alors

il vient, $\bar{\tilde{s}}^m_{ii'} = \frac{1}{2}(\tilde{s}^m_{ii} + \tilde{s}^m_{i'i'}) - \tilde{s}^m_{ii'}$. Cela nous amène à une matrice de dissimilarité $\overline{\tilde{S}}^m$. Les matrices $\tilde{S}$ et $\overline{\tilde{S}}$ sont alors obtenues en additionnant, respectivement, toutes les matrices $\tilde{S}^m$ et $\overline{\tilde{S}}^m$, soit $\tilde{S} = \sum_{m=1}^{M} \tilde{S}^m = \tilde{K}\tilde{K}^t$ et $\overline{\tilde{S}} = \sum_{m=1}^{M} \overline{\tilde{S}}^m$. La similarité globale entre deux objets O_i et $O_{i'}$ est donc $\tilde{s}_{ii'} = \sum_{m=1}^{M} \tilde{s}^m_{ii'}$ et leur dissimilarité globale est $\bar{\tilde{s}}_{ii'} = \sum_{m=1}^{M} \bar{\tilde{s}}^m_{ii'}$.

2.2 Maximisation du critère de Condorcet pondéré

Mathématiquement, le problème de la classification relationnelle de l'ensemble I en $\mathcal{K}$ classes disjointes se pose sous la forme d'un programme linéaire en variables bivalentes (0,1) : $\max_{X \in E(X)} \mathcal{R}_{AR}(\tilde{S},X)$ où

$$\mathcal{R}_{AR}(\tilde{S},X) = Tr(\tilde{S}X) + Tr(\overline{\tilde{S}}\overline{X}) = Tr[(\tilde{S} - \overline{\tilde{S}})X] + \beta \tag{1}$$

où $\beta = e^t\overline{\tilde{S}}e$ est une constante que l'on peut ignorer, car non pertinente d'un point de vue optimsation. X est la solution recherchée, elle modélise une partition dans un espace relationnel (une relation d'équivalence)., et doit vérifier les propriétés suivantes : $x_{ii} = 1, \forall i$ (réflexivité),$x_{ii'} - x_{i'i} = 0, \forall(i,i')$ (symétrie), $x_{ii'} + x_{i'i''} - x_{ii''} \leq 1, \forall(i,i',i'')$ (transitivité) et $x_{ii'} \in \{0,1\}, \forall(i,i')$ (binarité). $E(X)$ est l'ensemble des matrices carrées satisfaisant les propriétés d'une relation d'équivalence. $\overline{X}$ est la matrice complémentaire de X où chaque entrée $\bar{x}_{ii'} = 1 - x_{ii'}$.

3 Critère de Condorcet normalisé

Le critère de Condorcet pondéré considère le nombre de catégories (modalités) partagées par chaque paire d'objets, plus les modalités partagées par deux objets sont rares dans le jeu de données plus leur similarité est élevée. Néanmoins, ce critère n'est pas compensé par les cardinalités des classes. Ce qui signifie qu'une classe peut se faire petite quand elle est touchée par des valeurs aberrantes. Ainsi, nous définissons le nouveau critère que nous appelons critère de Condorcet normalisé et est donnée comme suit :

$$\tilde{\mathcal{R}}_{AR}(\tilde{S},X) = Tr[(\tilde{S} - \overline{\tilde{S}})V^{-1}X] \tag{2}$$

Où V est une matrice carrée diagonale de dimension $N \times N$ telle que $V = diag(Xe)$ et $v_{ii} = x_{i.}$ le nombre d'objets dans la même classe avec l'objet i. Selon le principe de Huygens, nous savons que l'inertie totale d'une partition X notée $I_T(X)$ est la somme de son inertie intra classe $I_W(X)$ et de son inertie inter classe $I_B(X)$, considérons ci-dessous la formule de Huygens : $\underbrace{\mathcal{I}_T(X)}_{InertieTotale} = \underbrace{\mathcal{I}_B(X)}_{InertieInterclasse} + \underbrace{\mathcal{I}_W(X)}_{InertieIntraclasse}$. En terme de l'opérateur matriciel Trace, ces inerties s'expriment :$\mathcal{I}_T = \frac{1}{N}Tr(\overline{\tilde{S}}\mathbf{1}_{N\times N}) = \frac{P}{M} - 1$, avec $\mathbf{1}_{N\times N} = ee^t$, $\mathcal{I}_B(X) = Tr(\tilde{S}V^{-1}X) - 1$ et $\mathcal{I}_W(\tilde{X}) = Tr(\overline{\tilde{S}}V^{-1}X)$. Nous pouvons observer que le critère de Condorcet normalisé s'exprime comme étant le critère de la différence inertielle $\tilde{\mathcal{R}}_{AR}(\tilde{S},X) = \mathcal{I}_B(X) - \mathcal{I}_W(X) + 1$. Suivant la formule de Huygens, on a les équivalences ci-aprés :

$$\max_X \tilde{\mathcal{R}}_{AR}(\tilde{S},X) \Leftrightarrow \max_X \mathcal{I}_B(X) \Leftrightarrow \min_X \mathcal{I}_W(X) \Leftrightarrow \max_X [\mathcal{I}_B(X) - \mathcal{I}_W(X)] \tag{3}$$

Il s'agit d'un résultat bien connu pour maximiser $\max_X \mathcal{I}_B(X)$ ou de réduire au minimum $\min_X \mathcal{I}_W(X)$.

4 Interprétation spectrale de la classification relationnelle

Trouver une partition de l'ensemble I qui maximise le critère de Condorcet normalisé est NP-difficile. Nous appliquons donc une relaxation spectrale du problème pour maximiser le critère de Condorcet normalisé.

4.1 Factorisation des relations X et $V^{-1}X$

Considérons la division des données en $\mathcal{K}$ classes disjointes où $\mathcal{K}$ est supérieur où égale à 2. Nous allons définir une matrice de partition Z de dimension $N \times \mathcal{K}$ avec une colonne pour chaque classe; $Z = (Z_1,Z_2,..,Z_k,..,Z_\mathcal{K})$. Chaque colonne est un vecteur d'indicatrices binaires, tels que $Z_{ik} = 1$(si l'objet i appartient à la classe k, 0 sinon). Notez que la somme de chaque ligne est égale à l'unité et la matrice Z satisfait les propriétés suivantes de $Tr(Z^T Z) = N$, et $Z^T Z = \mathcal{N} = diag(n_1,...,n_l,...,n_\mathcal{K})$ où n_k représente la cardinalité de la classe k. La relation d'équivalence X et la relation d'équivalence pondérée $V^{-1}X$ peuvent être maintenant factorisées de la manière suivante : $X = ZZ^t$ et $V^{-1}X = Z(Z^tZ)^{-1}Z^t$. La relation X étant une relation d'équivalence, elle pourra être également décomposée en un produit de trois matrices de la manière suivante : $X = \tilde{Z}\mathcal{N}\tilde{Z}^t$. La matrice $\tilde{Z}$ vérifie la propriété d'orthogonalité, en utilisant la décomposition de X on pourra réécrire le programme de classificaton relationnelle comme suit: $\max_{\tilde{Z}^t\tilde{Z}=I_\mathcal{K}} Tr[\mathcal{N}\tilde{Z}^t(\tilde{S} - \overline{\tilde{S}})\tilde{Z}]$, où $I_\mathcal{K}$ est la matrice identité d'ordre $\mathcal{K}$. Le programme ci-dessus est l'équivalent spectral du problème de maximisation du critère de Condorcet pondéré $\max_X \mathcal{R}_{AR}(\tilde{S},X)$, cette dernière formulation spectrale impose la connaissance a priori de la matrice $\mathcal{N}$, c'est à dire, le nombre de classes et la taille de chaque classe de la partition recherchée. Nous donnons maintenant l'équivalent spectral du problème de maximization du critère de Condorcet normalisé $\max_X \tilde{\mathcal{R}}_{AR}(\tilde{S},X)$: $\max_{\tilde{Z}^t\tilde{Z}=I_\mathcal{K}} Tr[\tilde{Z}^t(\tilde{S} - \overline{\tilde{S}})\tilde{Z}]$. Etant donné les équivalences présentées en (3), nous considérons dans la suite de ce papier le problème de maximisation de l'inertie inter classes $\mathcal{I}_B(X)$.

4.2 Algorithme SpectCat

Ci-aprés la procédure spectrale proposée, inspirée de l'algorithme proposé par Ng, Jordan et Weiss (Ng et al, 2001)

Algorithm2: SpectCat Algorithm

1. Construire la matrice de similarité $\tilde{S}$
2. Definir la matrice diagonale $D = diag(\tilde{S}e)$
3. Trouver la matrice U contenant les $\mathcal{K}$ premiers vecteur propres de la matrice $\hat{S} = D^{-1}\tilde{S}$
4. Construire la matrice normalisée $\hat{U}$ à partir de U : $\hat{U}_k = \frac{U_k}{||U_k||}$, $\forall k = 1,...,\mathcal{K}$
5. Considérons chaque ligne de $\hat{U}$ comme un point dans $R^\mathcal{K}$, Appliquer les k-means sur $\hat{U}$.
6. Affecter chaque objet i à la classe C_k si et seulement si la ligne correspondante $\hat{U}_i$ de $\hat{U}$ à été affectée à la classe C_k.

5 Expérimentation et validation

Une étude de performance a été réalisée pour évaluer notre méthode, dans cette section, nous décrivons les expériences et les résultats. Nous avons testé notre algorithme sur des données réelles obtenues à partir du référentiel UCI (Machine Learning Repository) et comparer ses performances avec d'autres algorithmes de clustering. Nous avons utilisé la pureté pour mesurer la qualité des résultats de clustering, nous effectuons des comparaisons sur sept bases de données UCI (voir tableau 1). Comme la méthode proposée est une approche spectrale adaptée

TAB. 1 – – *Description des bases de données*

Bases de données	# d'objects	# d'attributes	# de classes
Soybean small	47	21	4
Mushroom	8124	22	2
Congressional votes	435	16	2
Zoo	101	16	7
Hayes-roth	132	4	3
Balance Scale	625	4	3
Car evaluation	1728	6	4
Audiology	200	69	24

aux données catégorielles, nous avons comparé les performances de l'algorithme proposé avec d'autres algorithmes de classification de données catégorielles. Nous avons étudié la clustering trouvé par quatre algorithmes, notre algorithme SpectCat, k-modes standart, l'algorithme K-representative et l'algorithme Wk-modes (Aranganayagi et al, 2009). Du tableau 2, il est clair que la performance de la méthode proposée qui repose sur le principe de la classification spectrale donne des résultats meilleurs ou semblables à ceux d'autres approches, cela signifie que l'approche proposée améliore la pureté du clustering. les disparités des performances de SpectCat par rapport aux différents algorithmes peut s'expliquer par la structure interne aux données, à savoir le nombre de modalités par variable, l'effectif de chaque modalité et la cavité de la matrice de données.

TAB. 2 – – *mesure de Pureté pour K-modes, K-representatives, weighted k-modes et SpectCat*

Bases de données	K-Modes	K-Representatives	WK-Modes	SpectCat
Soybean small	66	89	89	100
Mushroom	59	61	61	61
Congressional votes	62	87	88	88
Zoo	88	89	90	90
Hayes-roth	41	42	42	54
Balance Scale	50	52	52	56
Car evaluation	70	70	71	70
Audiology	62	61	62	61

6 Conclusions et perspectives

Dans ce papier, nous avons étudié l'interprétation spectrale de la classification relationnelle des données catégorielles. Une procédure efficace pour l'optimisation est présentée, qui combine l'algorithme spectrale et la représentation relationnelle des données. Les résultats expérimentaux montrent l'efficacité de la méthode proposée. Nous avons l'intention de continuer notre travail par l'expérimentation d'un algorithme qui combine l'heuristique de l'AR avec celui de l'algorithme spectrale selon deux perspectives : la première repose sur l'utilisation de

l'AR comme une étape d'initialisation pour l'algorithme spectrale, dans la seconde perspective, nous allons proposer un crièrte hybride pour la classification des données mixtes provenant de deux sources différentes.

Références

Huang, Z. (1998). Extensions to the k-means algorithm for clustering large data sets with categorical values. Data Mining and Knowledge Discovery, 2, 283-304.

Newman, M. and Girvan, M.(2004). Finding and evaluating community structure in networks. Physical Review E, 69, 026113.

Li, T. Ma, S., and Ogihara, M. (2004). Entropy-based criterion in categorical clustering. Proceedings of The 2004 IEEE International Conference on Machine Learning (ICML 2004). 536-543.

Marcotorchino, J, F. (2006). *Relational analysis theory as a general approach to data analysis and data fusion*. In Cognitive Systems with interactive sensors, 2006.

Marcotorchino, J, F. Michaud, P. (1978). *Optimisation en analyse ordinale des données* In Masson, 1978.

White, S. and Smyth, P. (2005). A spectral clustering approach to finding communities in graphs. In SDM, pages 76-84.

Shi, J and Malik, J. (2000) "Normalized cuts and image segmentation," IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 22, no. 8, pp. 888-905, August 2000.

Ng, A. Y, Jordan, M, and Weiss Y, (2001) "On spectral clustering: Analysis and an algorithm," in Proc. of NIPS-14, 2001.

Aranganayagi, S and Thangavel. K, (2009) "Improved K-Modes for Categorical Clustering Using Weighted Dissimilarity Measure", International Journal of Engineering and Mathematical Sciences. vol5-2-19, 2009.

Summary

This paper introduces a spectral view on the Relational Analysis (RA) approach for categorical data clustering. It first, establishes the theoretical connection between the RA approach and classical spectral clustering technique. In particular, the RA problem is shown as a trace maximization problem, Thus usually this problem is converted by spectral relaxation into an optimization problem with a constraint which can be solved by Lagrange multipliers, and the solution is given by an eigenvalue problem.